Journal of Information Hiding and Multimedia Signal Processing ©2025 ISSN 2073-4212
Ubiquitous International Volume **16**, Number **3**, September **2025**

VPZL: Visual prompt-guided zero-shot learning for insulator defect detection

Pho Hai Dang

Faculty of Management Information Systems
Ho Chi Minh University of Banking
Ho Chi Minh City, Vietnam
dangph@hub.edu.vn

Received March 13, 2025, revised May 22, 2025, accepted May 24, 2025.

Abstract. *The combination of images and text enhances the capability of joint representation, facilitating tasks in image detection and recognition. In zero-shot learning (ZSL) models, representing extensive knowledge enables the model to generalize fundamental components, thereby predicting new labels without direct training. Notably, insulator defect detection encounters numerous challenges that necessitate leveraging both images and text to improve detection quality. However, these methods often struggle to fully grasp the context to focus on understanding features and enhancing reasoning ability for insulator defect detection. To address this, we propose a visual prompt-guided zero-shot learning system for insulator defect detection based on a similarity retrieval mechanism within the framework of narrowing the gap between semantic and visual features. Our method can suggest the most relevant attributes and objects for recognition based on visual and semantic features. Experiments on two standard benchmark datasets in both closed and open scenarios demonstrate promising results in supporting insulator defect detection.*
Keywords: Insulator defect detection, zero-shot learning, image processing

1. **Introduction.** Detecting Insulator defects [1] is a crucial task for safeguarding electrical power systems and ensuring timely maintenance and upkeep to prevent severe failures. However, acquiring data on Insulator defects is challenging due to the inherent hazards in power line environments. This necessitates the development of techniques capable of learning novel labels and attributes such as color, texture, and shape. The capacity to associate diverse attributes with various insulator defect instances is a facet of human perception, often referred to as feature binding, which aids in recognition and detection.

Zero-Shot Learning (ZSL) [2, 3] aims to develop training techniques for predicting novel labels and objects by leveraging attributes and objects that are not encountered or observed during the training phase. ZSL endeavors to concentrate on the reconstruction of known components to facilitate the recognition of new labels by efficiently enhancing features through integrated visual representations.

The training process seeks to augment knowledge by integrating computer vision with textual information and attempts to map concepts between images and semantic text. This integration enables tasks such as zero-shot classification and image recognition [4, 5, 6, 7]. However, integrating images and semantic text presents significant challenges. Furthermore, one of these components may suffer from data scarcity, such as a lack of image data or semantic text data. Consequently, this deficiency in generalization

capability can be attributed to the reliance on representational features of fixed classes, resulting in limited flexibility for novel classes.

Recent research in ZSL, focusing on the integration of images and semantic text for general recognition and detection problems and specifically for Insulator defect detection, remains relatively underexplored. Furthermore, general recognition and detection paradigms encounter limitations such as: (1) State-of-the-art methods often necessitate a fixed text format, for instance, "a photo of [object] with [attributes]," lacking true textual flexibility; (2) Algorithms are frequently limited to learning basic features like images or image bounding boxes for feature representation; (3) Predominant methods primarily concentrate on fixed data label sets, exhibiting inflexibility towards new label sets.

To address these challenges, we propose a Zero-Shot Learning (ZSL) based approach that integrates images and textual context for Insulator defect detection. Our main contributions are as follows:

- *Visual Guidance*: We introduce visual guidance by constructing embedding feature space sets to facilitate the learning of designs and the capture of visual patterns associated with attributes and objects.
- *Semantic Text Repository Development*: We develop a semantic text repository comprising several layers to support the recognition of novel classes in the future, as some classes may lack training images but semantic text can be proactively generated.
- *Insulator Defect Detection Adaptation*: Our proposed VPZL model incorporates an adaptation mechanism to dynamically adjust the training focus on label classes, thereby enhancing Insulator defect detection through the utilization of image features derived from a visual transformer.

2. **Related Works.** Zero-Shot Learning (ZSL) [2, 3] operates on the principle of recognizing untrained categories based on known foundational elements. Several ZSL approaches, such as leveraging distance-based methods [8], feature spaces [9], etc., aim to achieve structured generalization. Zabihzadeh et al. [10] developed a novel distance metric method for ZSL to facilitate object recognition. Visual feature embedding networks [11] are constructed to distinctly capture the diversity of patterns, attributes, and objects. Wang [12] proposed an attribute-embedding based approach to build attribute-based learning sets, thereby enhancing interaction when attributes change across different objects.

Wang and Yiheng [13] integrated images and semantic text to develop methods for attribute and object description to support object recognition and detection. Antwi [1] constructed a deep learning model for Insulator defect recognition and detection. Zhu [14] employed multi-modal deep learning for insulator defect detection to identify a wider range of Insulator defect instances. In general, methods relying on deep learning models utilizing neural networks still exhibit limitations in the accuracy of insulator defect detection.

Prompt learning based deep learning [15] leverages the attention mechanism of language models to enhance the quality of pattern recognition or object detection. Prompting models [16] endeavor to focus on the multi-modal nature of both textual and visual content. Huang et al. [17] combined ZSL with the creation of embedding spaces for textual and computer vision content to improve the quality of object recognition. With advancements in multi-modal learning techniques, prompt learning has seen new directions by focusing on the multi-modality of both textual and visual content to construct embedding feature spaces [15].

3. **Proposed Method.**

3.1. **Problem Definition.** Ensuring the reliability and safety of electrical power systems necessitates the automated detection of insulator defects, a task of paramount importance. Traditional supervised learning approaches for object detection require extensive labeled data for each defect type, rendering them impractical for insulator defects. This impracticality stems from the rarity of certain defect types and the inherent challenges in acquiring comprehensive datasets under hazardous power line conditions. To address these limitations, we explore a ZSL paradigm for insulator defect detection, capitalizing on the compositional nature of defects and insulators.

Formally, let $A = \{a_0, a_1, ..., a_n\}$ denote the set of attributes relevant to describing insulators and their defects, encompassing visual characteristics such as color, texture, and shape, as well as semantic descriptors. Let $O = \{o_0, o_1, ..., o_m\}$ represent the set of objects, which in our context primarily includes objects related to insulators and potentially distinct categories of defects. We define the composition space C as the Cartesian product of attributes and objects, $C = A \times O$. Each element within C embodies a plausible description of an insulator or an insulator defect, derived from a combination of attributes and objects.

We acknowledge that the set of all possible compositions C can be partitioned into two disjoint subsets: seen compositions C_s encountered during training, and unseen compositions C_u absent from the training data. Our objective is to develop a model capable of detecting and classifying insulator defects, particularly those belonging to the unseen composition set C_u, effectively performing Zero-Shot Learning.

We further delineate two distinct scenarios within Compositional Zero-Shot Learning (CZSL):

- Closed-World CZSL: In this setting, we operate under the premise that all possible compositions encountered during testing are drawn from a predefined subset $C_{test} \subseteq C$. While the model may not have been exposed to all compositions during training, we assume that the test compositions reside within a known, bounded space of possibilities, $C_{test} = C_s \cup C_u$. The task is to learn a function $f : X \to C_{test}$, where X represents the input space of insulator images. The model must classify an input image into one of the compositions within this predefined C_{test}, encompassing both seen and unseen combinations during training but confined to this known set.
- Open-World CZSL: In a more demanding Open-World scenario, we relax the assumption of a predefined test composition set. The model is expected to navigate the entire composition space $C = A \times O$, encompassing both feasible and potentially infeasible combinations. This implies that the model may encounter entirely novel compositions during testing, compositions neither seen during training nor explicitly anticipated within a predefined test set. The objective in this case is to learn a function $f : X \to C$, requiring the model to generalize to a truly open-ended space of insulator defect compositions.

For both Closed-World and Open-World CZSL settings, we aim to harness the synergy of image and text encoders. The image encoder will extract visual features from insulator images, while the text encoder will process semantic information pertaining to attributes and objects. By effectively bridging the gap between visual and semantic spaces, we aspire to empower the model to recognize and detect insulator defects, even for novel compositions unseen during training. This approach is particularly pertinent to real-world insulator defect detection, where the variability in defect types and appearances necessitates robust generalization capabilities that transcend traditional supervised learning methodologies.

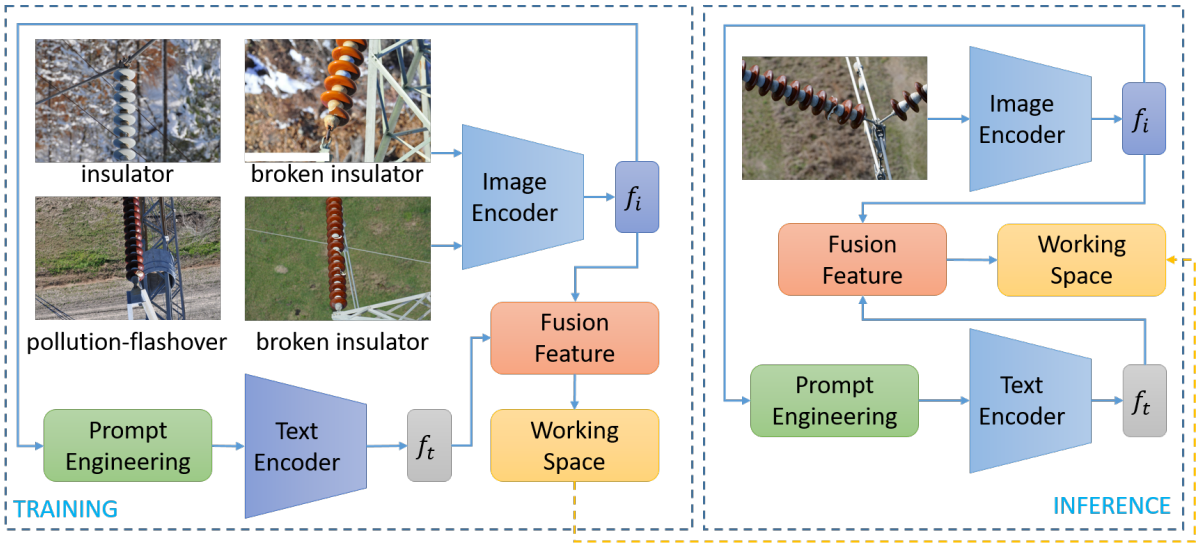

FIGURE 1. The framework of the proposed VPZL

3.2. **Model Architecture.** Figure 1 illustrates the proposed model architecture, designed to leverage both visual and textual information to enhance object detection capabilities, particularly in the context of insulator defect detection. This architecture employs a dual encoding framework, processing image and text input data in parallel before merging their representations in a shared workspace. This multimodal approach aims to exploit the complementary strengths of visual and semantic signals to improve detection performance, especially for tasks such as ZSL.

The image processing branch, depicted in the upper portion of the architecture, specializes in processing image input data. As illustrated, the example images show various insulator conditions, including 'insulator' (normal), 'broken insulator,' and 'surface discharge due to pollution,' demonstrating the model's intended application in detecting various insulator states, including defects. Raw image data, representing insulators and their surrounding environment, is ingested into the system. These images are designed to capture the visual characteristics of insulators, which are crucial for identifying potential defects. The image input data is then fed into an 'Image Encoder.' This module is responsible for extracting salient visual features from the input images. The specific type of Image Encoder is specified in the diagram, including Convolutional Neural Networks (CNNs), Vision Transformers (ViTs) [18], and other suitable architectures capable of learning hierarchical and discriminative visual representations. The output of the Image Encoder is a feature vector denoted as f_i. This vector represents the encoded visual features of the input image in a high-dimensional space.

The text processing branch, located in the lower portion of the architecture, is designed to process textual information. This branch highlights the incorporation of 'Prompt Engineering' and a 'Text Encoder' to effectively utilize textual signals. The 'Prompt Engineering' component suggests that the system uses prompts to guide the text encoder in extracting relevant semantic information. 'Prompt Engineering' may involve constructing specific textual prompts that provide contextual information or descriptions related to objects or attributes of interest (e.g., 'detect broken insulator,' 'identify surface discharge due to pollution'). These prompts, along with relevant textual data, serve as input to the Text Encoder. The 'Text Encoder' processes the engineered prompts and related textual data. This module is responsible for transforming the textual input data into a semantic

embedding. Similar to the Image Encoder, the specific type of Text Encoder is specified, including Transformer-based models like BERT [19] capable of capturing semantic relationships and contextual information from text. The output of the Text Encoder is a feature vector denoted as f_t. This vector represents the encoded semantic features derived from the textual input data.

After separately encoding the image and text input data, the architecture proceeds to merge these representations to create a joint multimodal representation. The 'Feature Fusion' module receives the image feature vector f_i and text feature vector f_t as inputs. This module is responsible for integrating visual and semantic information. The diagram does not specify the fusion mechanism, which may include techniques such as concatenation, element-wise operations (addition, multiplication), attention mechanisms, or more complex fusion networks. The goal of this module is to produce a combined feature representation that effectively captures the relationships and interactions between visual and textual signals. The output of the 'Feature Fusion' module leads to a 'Workspace.' This space can be understood as a joint embedding space where the fused visual and textual features are represented. In this workspace, the model can perform downstream tasks such as object detection, classification, or recognition, leveraging the enriched multimodal representation. The workspace allows the model to compare and associate visual and textual features for effective object recognition and detection, particularly in scenarios requiring semantic understanding, such as Zero-Shot Learning, where textual descriptions can generalize to unseen object categories.

The described architecture represents a multimodal approach that effectively integrates visual and textual information for object detection. By processing images and text in parallel and then fusing their representations, the model aims to achieve more robust and semantically informed detection capabilities. The inclusion of Prompt Engineering [21] highlights the importance of guiding the text encoder with relevant prompts, indicating the intention to enhance the model's ability to understand and utilize textual descriptions for improved object recognition, particularly in tasks requiring generalization to new categories or attributes as commonly found in Zero-Shot Learning scenarios. This architecture is particularly suitable for complex image recognition tasks, such as insulator defect detection, where contextual and semantic information can significantly improve performance.

3.3. **Loss function.** To effectively train the proposed model architecture for ZSL based insulator defect detection, we employ a contrastive loss function designed to align the visual and textual feature representations within the shared working space. The primary objective of this loss function is to ensure that semantically related image and text embeddings are mapped to close proximity in the joint embedding space, while embeddings from unrelated pairs are pushed further apart. This mechanism facilitates the model's ability to generalize to unseen categories by understanding the underlying relationships between visual features and semantic descriptions.

Let us denote the image feature vector extracted from the Image Encoder as f_i and the text feature vector extracted from the Text Encoder as f_t. We define a similarity function $S(f_i, f_t)$ to measure the compatibility between the image and text embeddings. A commonly used similarity measure is the cosine similarity, defined as:

$$S(f_i, f_t) = \frac{f_i \cdot f_t}{\|f_i\|\|f_t\|} \tag{1}$$

where $\cdot$ represents the dot product, and $\|\cdot\|$ denotes the Euclidean norm.

Our loss function is formulated as a margin-based contrastive loss. For each training instance, we consider positive pairs, which consist of an image and its corresponding text description (e.g., an image of a "broken insulator" and the text prompt "broken insulator"), and negative pairs, which are composed of an image and a mismatched text description (e.g., an image of a "normal insulator" and the text prompt "broken insulator"). The loss function is then defined as:

$$L = \sum_{(i,t)^+} \sum_{(i,t)^-} \max(0, m - S(f_i^+, f_t^+) + S(f_i^+, f_t^-)) \tag{2}$$

where:

- $(i,t)^+$ represents a positive pair of image and text embeddings, with f_i^+ and f_t^+ being their respective feature vectors.
- $(i,t)^-$ represents a negative pair, with f_i^+ being the image feature vector from the positive pair, and f_t^- being the text feature vector from a mismatched text description.
- m is a margin parameter, typically set to a positive value (e.g., $m = 1$), which controls the desired separation between positive and negative pairs.

The loss function L is minimized during training. Minimizing L encourages the similarity $S(f_i^+, f_t^+)$ between positive pairs to be high (ideally close to 1 for cosine similarity), and the similarity $S(f_i^+, f_t^-)$ between negative pairs to be low (ideally less than $m - S(f_i^+, f_t^+)$). In essence, the model is penalized when the similarity of a negative pair is not sufficiently smaller than that of a positive pair by at least the margin m.

By optimizing this contrastive loss, we aim to learn an embedding space where visual features of insulator defects are closely aligned with their corresponding semantic descriptions. This alignment is crucial for enabling the model to effectively perform Zero-Shot Learning, allowing it to detect and classify novel insulator defect types based on learned semantic relationships, even for categories not explicitly seen during training.

4. **Experiments.**

4.1. **Dataset.**

To comprehensively evaluate the proposed model's performance and generalization capabilities, we employ two distinct datasets: Ins-States and MIT-States [20]. The Ins-States dataset is a specialized dataset for insulator defect detection that we have collected, while the MIT-States dataset is a publicly available compositional dataset for broader evaluation.

The Ins-States dataset is specifically constructed for the task of insulator defect detection and is organized into three distinct subsets to facilitate model development and evaluation. As depicted in the data card, the dataset is structured into `train`, `val`, and `test` directories. The training set (`train`) comprises the largest portion with 1296 images, intended for model parameter optimization. The validation set (`val`), consisting of 144 images, is utilized for model validation and hyperparameter tuning during the training process. Finally, the test set (`test`), containing 160 images, is reserved for assessing the final performance of the trained model. The Ins-States dataset focuses on capturing a variety of insulator conditions, encompassing both normal and defective states, rendering it a highly relevant resource for our insulator defect detection task. Within the Ins-States dataset, we designate the following classes as "seen" during the training phase: (1) Normal Insulator: This fundamental class represents insulators in their operational, defect-free state. It is essential for the model to learn baseline insulator characteristics and to differentiate them from defective conditions. (2) Seen Insulator Defect Types: Depending on the diversity of defect types within the Ins-States dataset, a selection of the most

prevalent defect types may be designated as "seen" classes. For instance, if the dataset encompasses defect types such as chipped insulators, surface cracks, or minor corrosion, these could be included as "seen" classes. Conversely, the following classes are designated as "unseen" and are reserved exclusively for ZSL testing purposes: (i) Broken Insulator: This class represents insulators exhibiting severe physical damage, often resulting from physical impact or aging processes. "Broken insulators" are a critical defect type in practical applications; however, this class is intentionally kept "unseen" during training to evaluate the ZSL capability of the model; (ii) Pollution-Flashover: This class describes surface flashover events caused by the accumulation of pollutants on the insulator surface. "Pollution-flashover" represents a distinct defect type compared to "broken insulators" and is also retained within the "unseen" set to assess the model's ability to generalize to novel defect categories; (iii) Other Unseen Insulator Defect Types: In addition to "broken insulator" and "pollution-flashover", should the Ins-States dataset contain further, less prevalent or data-scarce defect types (e.g., punctured insulators, deep cracks, etc.), these could also be incorporated into the "unseen" set to enhance the challenge of the ZSL task.

In addition to the domain-specific Ins-States dataset, we leverage the MIT-States dataset to evaluate the generalization capabilities of our model on a more general compositional recognition task. MIT-States, a publicly accessible dataset gathered from web-crawled images, is characterized by its rich compositional structure. It encompasses a diverse range of 245 distinct object categories and 115 diverse attribute categories. The MIT-States dataset serves as a challenging benchmark for evaluating compositional generalization, enabling us to examine the model's capacity to handle a wider spectrum of visual concepts beyond the insulator defect domain.

By utilizing both the Ins-States dataset and the MIT-States dataset, we aim to achieve a rigorous evaluation of our proposed approach. The Ins-States dataset provides a focused assessment on the target task of insulator defect detection, while the MIT-States dataset allows us to gauge the broader generalization abilities of our model in a compositional zero-shot learning context.

4.2. **Experiment setup.** The VPZL model is implemented using the PyTorch framework. Model optimization is performed using the Adam optimizer across the Ins-States and MIT-States datasets. For feature extraction, we leverage pretrained models for both image and text encoding. Specifically, the image encoder is initialized with a pretrained Vision Transformer (ViT) model, while the text encoder is based on a pretrained BERT model. All experiments are conducted on four NVIDIA RTX 4090 GPUs, ensuring consistent computational resources across evaluations. The batch size, denoted as M, is set to 32 for the Ins-States dataset and 30 for the MIT-States dataset. The smaller batch size for MIT-States is attributed to the dataset's greater compositional diversity, which necessitates adjustments for efficient training.

Our experimental evaluation is designed to address the following key research questions:

- RQ1: How does the VPZL model perform compared to state-of-the-art methods?
- RQ2: How does the VPZL model predict in practice to detect insulator defects?

4.3. **Performance Compare (RQ1).** Table 1 presents the performance comparison of different methods, including TMN, SymNet, CGE, CompCos, and the proposed VPZL model, in an Open World environment on both MIT-States and Ins-States datasets. The evaluation metrics include accuracy on seen compositions ('Seen'), accuracy on unseen compositions ('Unseen'), and Area Under the Curve (AUC). Overall, the VPZL model significantly outperforms the comparison methods across both datasets and all metrics, achieving the highest scores, highlighted in bold in the table. This is particularly crucial

TABLE 1. Open-World Results on MIT-States and Ins-States. The results are reported for Seen, Unseen, and Area Under the Curve (AUC). Bold and blue indicate the first and second best results, respectively

Method	MIT-States			Ins-States		
	Seen	**Unseen**	**AUC**	**Seen**	**Unseen**	**AUC**
TMN [25]	12.6	0.9	0.1	56.4	42.6	19.2
SymNet [23]	21.4	7.0	0.8	54.8	43.1	18.5
CGE [24]	32.4	5.1	1.0	62.7	47.3	22.9
CompCos [25]	25.4	10.0	1.6	59.5	46.5	21.3
VPZL (Ours)	**30.5**	**15.9**	**3.3**	**66.2**	**60.0**	**30.8**

in the context of insulator defect detection, where the ability to recognize new defect types not encountered during training ('Unseen' accuracy) is vital.

On the MIT-States dataset, VPZL demonstrates a 'Seen' accuracy of 30.5% and an 'Unseen' accuracy of 15.9%, achieving an AUC of 3.3. Although MIT-States is a more general dataset, these results show VPZL's superiority, particularly in 'Unseen' accuracy, which is a major challenge in ZSL. Compared to other methods on MIT-States, CompCos achieves the second-best 'Unseen' accuracy at 10.0%, and CGE achieves the second-best 'Seen' accuracy at 32.4%, though CGE's 'Unseen' performance is significantly lower than VPZL's.

More importantly, on the Ins-States dataset, specialized for insulator defect detection, VPZL continues to demonstrate superior performance with a 'Seen' accuracy of 66.2%, an impressive 'Unseen' accuracy of 60.0%, and an AUC of 30.8. Notably, while CGE and CompCos show relatively competitive 'Seen' accuracy on Ins-States, their 'Unseen' accuracy and AUC scores remain significantly lower than VPZL's. In the context of insulator defect detection, VPZL's high 'Unseen' accuracy on Ins-States is particularly significant. It demonstrates the model's strong generalization capability to new, previously untrained insulator defect types. In practical applications, insulator defect detection systems often face a wide variety of defects, many of which may be new variants or rare cases. Therefore, VPZL's robust ZSL capability, evidenced by its high 'Unseen' accuracy, provides a significant advantage in real-world deployment.

These results emphasize the effectiveness of the VPZL method in open-world compositional zero-shot learning, particularly in the insulator defect detection problem. VPZL's enhanced generalization capability to novel attribute-object combinations, clearly demonstrated on the challenging Ins-States dataset, proves the advantage of the model architecture in handling open-world scenario complexities and effectively learning transferable representations for unseen compositions. The significant performance gap compared to other methods further reinforces VPZL's advantage in practical insulator defect detection applications, where recognizing new and diverse defect types is crucial.

4.4. **Qualitative Study (RQ2).** Figure 2 shows the prediction results of the insulator defect detection model in two scenarios: Closed World and Open World. In each scenario, the image displays three different examples of insulators, accompanied by ground truth (GT) labels - actual labels assigned by humans - and the top three predictions made by the model. This analysis highlights the differences in model performance and behavior when faced with different assumptions about the test label space, which is crucial in the context of Zero-Shot Learning (ZSL) and particularly in the insulator defect detection problem.

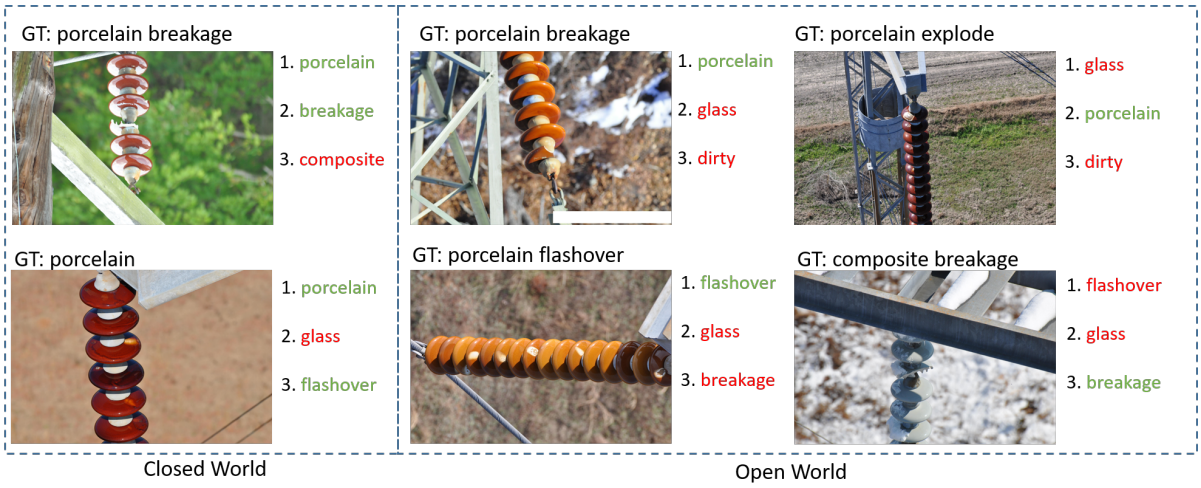

FIGURE 2. The detection results of VPZL model.

In the Closed World scenario, the model operates within a restricted and predefined label space. In the first example (GT: porcelain breakage), the model's top three predictions all focus on material attributes ('porcelain', 'composite') and defect type ('breakage'). Similarly, in the second example (GT: porcelain), the predictions also revolve around materials ('porcelain', 'glass') and surface condition ('flashover'). This demonstrates that in the Closed World, the model tends to predict familiar labels within the predefined label space, focusing on accurate classification within known labels.

Conversely, in the Open World scenario, the model must deal with a broader label space that may contain labels unseen during training. This is clearly illustrated through the examples in the figure. In the third example (GT: porcelain explode), while the top prediction still relates to material ('glass', 'porcelain'), the third prediction shows the 'dirty' label, a surface condition attribute, indicating that the model begins to explore broader descriptive aspects when facing an open label space. Similarly, in the fourth example (GT: porcelain flashover), the 'flashover' label is correctly predicted in the first position, showing the ability to recognize the target defect. However, subsequent predictions expand to other material attributes and defect types ('glass', 'breakage'). Finally, in the fifth example (GT: composite breakage), the top prediction continues to be 'flashover', and subsequent predictions relate to different materials and defect types ('glass', 'breakage').

This analysis demonstrates that the VPZL model, trained using the Zero-Shot Learning approach, shows effective operational capability in both Closed World and Open World scenarios. In the Closed World, the model focuses on accurate classification within the known label space. Meanwhile, in the Open World, the model tends to explore and propose more diverse labels, potentially including new attributes and defect types, demonstrating its ability to generalize and adapt to an open label space. This is particularly important in the insulator defect detection problem, where various types of defects may appear, including rare or previously unknown types during training. The ability to perform well in both scenarios, especially the generalization capability in the Open World, confirms the superiority and practical application potential of the VPZL model in insulator defect detection.

5. **Conclusions.**

To comprehensively evaluate the VPZL model, we constructed and utilized the Ins-States dataset specifically for insulator defect detection, alongside the public MIT-States dataset to test general generalization capability. Experiments were conducted

in both Closed World and Open World settings, using a strict Zero-Shot Learning protocol where the model is trained on a set of 'seen' classes and evaluated on 'unseen' classes. Experimental results on both datasets, particularly on the Ins-States dataset, convincingly demonstrated the superiority of the VPZL model compared to other state-of-the-art methods such as TMN, SymNet, CGE, and CompCos. VPZL showed significant improvements in accuracy on 'unseen' classes and AUC scores, indicating strong generalization capability to new, previously unseen insulator defect types. Qualitative analysis of prediction results also highlighted VPZL's ability to adapt to both restricted label space (Closed World) and open label space (Open World), demonstrating high flexibility and practical applicability.

The main contributions of this research include proposing the unique VPZL architecture, constructing the specialized Ins-States dataset, and experimentally demonstrating VPZL's effectiveness in the Zero-Shot Learning approach to insulator defect detection. These research findings open up a new and promising direction for developing intelligent insulator defect detection systems capable of operating effectively in real-world environments, where the diversity and novelty of defect types pose significant challenges. In the future, the research could be extended to explore deeper aspects such as the interpretability of VPZL's predictions, optimization of architecture and loss functions for further performance improvement, and application of the model in large-scale real-world insulator defect detection systems.

Acknowledgment.

REFERENCES

[1] Antwi-Bekoe, Eldad, et al. "A deep learning approach for insulator instance segmentation and defect detection." Neural Computing and Applications 34.9 (2022): 7253-7269.

[2] Lazaros, Konstantinos, et al. "A comprehensive review on zero-shot-learning techniques." Intelligent Decision Technologies 18.2 (2024): 1001-1028.

[3] Gull, Muqaddas, and Omar Arif. "Multi-Label Zero-Shot Learning with Adversarial and Variational Techniques." IEEE Access (2024).

[4] Shermin, Tasfia, et al. "Integrated generalized zero-shot learning for fine-grained classification." Pattern Recognition 122 (2022): 108246.

[5] Tran-Anh, Dat, et al. "Integrative zero-shot learning for fruit recognition." Multimedia Tools and Applications 83.29 (2024): 73191-73213.

[6] Long Duong Phi*, Binh Pham Nguyen Thanh, and Quang Tran Van A Classification Method based on Concatenation Features for Diagnosing Skin Diseases Journal of Information Hiding and Multimedia Signal Processing, Vol. 16, No. 1, pp. 401-413, March 2025.

[7] Budi Setiyono, Dwi Ratna Sulistyaningrum, Ario Fajar Pratama, Ridho Nur Rohman Wijaya A Modification of the Temporal Group Attention Method on Super-Resolution Video for Vehicle Number Plate Detection Journal of Information Hiding and Multimedia Signal Processing, Vol. 15, No. 3, pp. 156-165, September 2024.

[8] Xue, Dinghao, et al. "Leveraging Foundation Models for Zero-Shot IoT Sensing." arXiv preprint arXiv:2407.19893 (2024).

[9] Kong, Xia, et al. "En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[10] Zabihzadeh, Davood, and Mina Masoudifar. "ZS-DML: Zero-Shot Deep Metric Learning approach for plant leaf disease classification." Multimedia Tools and Applications 83.18 (2024): 54147-54164.

[11] Behera, Sushree S., Dwarikanath Mahapatra, and Sudipta Roy. "Semantics-Guided Label Synthesizer for Generalized Zero-Shot Object Classification." Available at SSRN 5106958.

[12] Wang, Chengji, et al. "Improving embedding learning by virtual attribute decoupling for text-based person search." Neural Computing and Applications (2022): 1-23.

[13] Wang, Yiheng, et al. "Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning." Advanced Engineering Informatics 53 (2022): 101699.
[14] Zhu, Bingqian, et al. "Enhanced YOLOv8 Network Utilizing RGB-D Multi-Modal and Multi-Scale Feature Fusion for Defect Detection in Power Transmission Scenes." 2024 4th International Conference on Energy Engineering and Power Systems (EEPS). IEEE, 2024.
[15] Li, Guilin, et al. "Multimodal Inplace Prompt Tuning for Open-set Object Detection." Proceedings of the 32nd ACM International Conference on Multimedia. 2024.
[16] Yang, Lingfeng, et al. "Fine-Grained Visual Text Prompting." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).
[17] Huang, Kaiqiang, Luis Miralles-Pechuán, and Susan Mckeever. "Enhancing zero-shot action recognition in videos by combining GANs with text and images." SN Computer Science 4.4 (2023): 375.
[18] Gehrig, Mathias, and Davide Scaramuzza. "Recurrent vision transformers for object detection with event cameras." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.
[19] Dong, Xiaoyi, et al. "Peco: Perceptual codebook for bert pre-training of vision transformers." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 1. 2023.
[20] Saini, Nirat, Khoi Pham, and Abhinav Shrivastava. "Disentangling visual embeddings for attributes and objects." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[21] Zhou, Kaiyang, et al. "Conditional prompt learning for vision-language models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[22] Mancini, Massimiliano, et al. "Open world compositional zero-shot learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[23] Anwaar, Muhammad Umer, et al. "A contrastive learning approach for compositional zero-shot learning." Proceedings of the 2021 International Conference on Multimodal Interaction. 2021.
[24] Naeem, Muhammad Ferjad, et al. "Learning graph embeddings for compositional zero-shot learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[25] Mancini, Massimiliano, et al. "Open world compositional zero-shot learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.